Introduction

The main goal of project is , “**How can we develop a system that accurately converts handwritten text into digital format while effectively addressing the challenges posed by the diversity and variability of individual handwriting styles?**”. The project can be further broken down in to sub-problems such as:

- **Handwritten Text Recognition:** Develop a model capable of accurately recognizing handwritten characters and words.
- **Handwritten Text Segmentation:** Implement effective techniques to segment handwritten text into individual characters or words, addressing overlapping characters and inconsistent spacing.
- **Diversity and Variability in Handwriting:** Address the challenges posed by the diversity and variability in handwriting styles to improve recognition accuracy through robust preprocessing techniques and adaptive learning algorithms.

General Review

The literature review on handwritten document conversion revealed that the process can be broadly divided into two main tasks: handwriting detection/segmentation and handwriting recognition.

1. **Handwriting Detection/Segmentation**
    - **Techniques**: Computer vision tricks, transformers, and layout parsers.
    - **Purpose**: To effectively segment handwritten text into individual characters or words for further processing.
2. **Handwriting Recognition**
    - **Techniques**: Transformers and CRNN models (Convolutional Recurrent Neural Networks).
    - **Purpose**: To accurately recognize and transcribe the segmented handwritten characters or words into digital text.

Datasets Available

1. **IAM Dataset**: A widely used dataset containing forms of handwritten English text.
2. **Sagemaker Dataset**: Another dataset available for handwriting recognition tasks.
3. **Crowd-Sourced Dataset**: A dataset obtained through crowd-sourcing efforts, providing diverse handwriting samples.

Common Models Used

1. **CNN (Convolutional Neural Networks)**: Commonly used for image-based tasks, including handwriting recognition.
2. **RNN (Recurrent Neural Networks)**: Effective for sequence prediction tasks, including text recognition.
3. **CRNN (Convolutional Recurrent Neural Networks)**: Combines the strengths of CNNs and RNNs for efficient handwriting recognition.

Major Issues Identified

1. **Variability in Handwriting Styles**: Handwriting differs significantly among individuals, posing a challenge for recognition models.
2. **Segmentation Accuracy**: Accurate segmentation of handwritten text is crucial for effective recognition, yet it remains challenging due to overlapping characters and inconsistent spacing.
3. **Dataset Limitations**: Existing datasets may not cover the full range of handwriting variability, requiring more diverse and comprehensive data for training robust models.

Paper Review

TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models

DOI: https://doi.org/10.48550/arXiv.2109.10282

Main Idea

The paper introduced TrOCR, a Transformer-based Optical Character Recognition (OCR) model that uses pre-trained models. The main contribution is an end-to-end OCR framework that significantly improves recognition performance by using a Transformer architecture.

Datasets Used

- IAM Dataset
- Synthesized handwritten text line images by the TRDG2 , an open-source text recognition data generator
- IIIT-HWS dataset
- SROIE dataset

Code Base

https://github.com/microsoft/unilm/tree/master/trocr

Model(s) Used

Transformer architecture

Procedure/ Configuration

1. Model Initialization: Pre-trained weights from large-scale datasets are used to initialize transformer layers
2. Pre-Training: Model is pre-trained on general text recognition tasks using large datastes to learn text features.
3. Fine-Tuning: The pre-trained model is fine tuned on OCR-specific datasets like IAM for better performance on handwritten text
4. Task Pipeline: Images are preprocessed(resized and normalized), passed through the encoder, and the decoder generates text sequences.
5. Data Augmentation: Techniques such as rotations, scaling, cropping and adding noise are used to augment training data.

Performance Metrics Used

- Precision
- Recall
- F1
- CER

SROIE DATASET	TrOCR(small)	TrOCR(base)	TrOCR(large)
Recall	**95.89**	**96.37**	**96.59**
Precision	**95.74**	**96.31**	**96.57**
F1	**95.82**	**96.34**	**96.58**
IAM Dataset + Synthetic Dataset	**TrOCR(small)**	**TrOCR(base)**	**TrOCR(large)**
CER	**4.22**	**3.42**	**2.89**

Key Learning/Finding

1. Effectiveness of Transformers: The TrOCR model showcases the effectiveness of using Transformer architectures for OCR tasks, achieving significant improvements in text recognition accuracy.
2. Pre-training Benifits: Utilizing pre-trained models helps in leveraging large-scale datasets, which enhance the model’s ability to generalize from limited OCR-specific data.

An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition

DOI: 10.1109/TPAMI.2016.2646371

Main Idea

This paper introduces a novel end-to-end trainable neural network for scene text recognition, integrating feature extraction, sequence modeling, and transcription into a unified framework. The research addresses the challenges of recognizing text in natural scene images, focusing on handling sequences of arbitrary length without requiring character segmentaion or scale normalization. The contributions include improved performance on both lexicon-free and lexicon-based tasks, and the development of a practical, smaller model for real-world applications.

Datasets Used:

- IIIT-5k
- Street View Text
- ICDAR 2003
- ICDAR 2013

Model(s) Used

Convolutional Recurrent Neural Network (CRNN)

Code Base Link:
https://github.com/bgshih/crnn

Procedure/Configuration or Test:

1. Model Architecture: CRNN combines convolutional layers for feature extraction, recurrent layers for sequence modeling, and a transcription layer for final recognition.
2. Initialization: Parameters are initialized using standard techniques like Xavier initialization.
3. Task Pipeline: End-to-end training pipeline without the need for character segmentation.

Particular Techniques/Tricks Applied:

1. End-to-end training to simplify the pipeline and improve accuracy.

2. Lexicon-free and lexicon-based recognition to handle various real-world scenarios.

Performance Metrics Used:

Accuracy on standard benchmarks:

- ICDAR 2013: 86.7%
- SVT: 80.8%

Key Learning/Finding:

The CRNN model demonstrates the effectiveness of an integrated approach for scene text recognition, handling sequences of arbitrary lengths and achieving high accuracy on standard benchmarks. The end-to-end trainable framework simplifies the recognition pipeline and enhances the model's practicality for real-world applications.

Full Page Handwriting Recognition via Image to Sequence Extraction

DOI: 10.1109/ACCESS.2018.2810892

Main Idea and Contribution

The paper presents a neural network-based Handwritten Text Recognition (HTR) model that can recognize full pages of handwritten or printed text without segmentation. The model can handle various text orientations, layouts, and sizes, and can generate auxiliary markup related to formatting and layout. The key contribution is an end-to-end model architecture combining a ResNet encoder and a Transformer decoder, achieving state-of-the-art performance on the IAM dataset and surpassing commercial HTR APIs on the proprietary Free Form Answers dataset.

Dataset(s) Used

- IAM Dataset
- Free Form Answers Dataset: Proprietary dataset derived from scans of test paper submissions in STEM subjects, including text, math equations, tables, drawings, and diagrams.

Model(s) Used

- Encoder: ResNet (without the last two layers: average-pool and linear projection).
- Decoder: Transformer, using attention-based sequence-to-sequence architecture.

Procedure/Configuration or Test

The model architecture consists of:

1. A ResNet-based encoder extracting a 2D feature-map from the input image.
2. A Transformer-based decoder converting the encoded representation into text.

Key configurations:

1. 2D positional encoding added to the feature-map.
2. Transformer decoder with causal self-attention and encoder-decoder attention.
3. Character-level vocabulary for transcription, with auxiliary markup tags for non-text regions.

Performance Metrics

CER

1. IAM Dataset = 6.3%
2. Free Form Answers = 7.6%
3. Wikitext (1 column) 0.008%

4. Wikitext (2 column) = 0.012%

Key Learning/Findings

- End-to-end models can significantly simplify and improve the process of full-page handwriting recognition by eliminating the need for prior segmentation and manual preprocessing.
- Combining CNN and Transformer architectures leverages the strengths of both image and language processing models, enabling robust and flexible HTR systems.
- Character-level transcription with auxiliary markup allows the model to handle diverse and complex layouts without specific adaptations.

Decoupled Attention Network for Text Recognition

DOI: 10.1109/ICCV.2019.00066

Main Idea and Contribution

The paper proposes a Decoupled Attention Network (DAN) for text recognition, aiming to address the alignment problems in traditional attention mechanisms. The main contributions are:

1. Convolutional Alignment Module (CAM): This replaces the traditional recurrency alignment module, performing alignment using only visual information without relying on historical decoding information, thus eliminating misalignment caused by decoding errors.
2. Decoupled Attention Network (DAN): An effective, flexible, and robust end-to-end text recognizer that consists of a feature encoder, a convolutional alignment module, and a decoupled text decoder.
3. Performance: DAN achieves state-of-the-art performance on multiple text recognition tasks, including offline handwritten text recognition and regular/irregular scene text recognition.

Dataset(s) Used

- IAM Dataset
- RIMES

Code Base

https://github.com/Wang-Tianwei/Decoupled-attention-network

Model(s) Used

- **Feature Encoder**: Based on a convolutional neural network (CNN) to extract visual features from input images.
- **Convolutional Alignment Module (CAM)**: Uses a fully convolutional network (FCN) in a channel-wise manner to generate attention maps.
- **Decoupled Text Decoder**: Uses the feature map and attention maps for final prediction with a gated recurrent unit (GRU).

Procedure/Configuration or Test

- Feature Encoder Configuration: The encoder consists of several residual blocks with specific configurations detailed in the paper.

- CAM: Performs attention operation in a channel-wise manner, which is different from current attention mechanisms. It is flexible and can switch between 1D and 2D forms by adjusting the downsampling ratio.
- Testing Strategy: Evaluated on tasks like handwritten text recognition and scene text recognition to demonstrate effectiveness, flexibility, and robustness.

Performance Metrics

CER

1. IAM Dataset = 6.4%
2. RIMES = 2.7%

WER

IAM Dataset = 19.6%

RIMES = 8.9%

Key Learning/Findings

- **Alignment Independence**: Decoupling alignment from decoding significantly improves the robustness of the text recognizer, reducing the impact of decoding errors on alignment.
- **Versatility**: The ability to switch between 1D and 2D forms makes DAN adaptable to different text recognition scenarios, from handwritten to scene texts.
- **Simplified Attention Mechanism**: The use of a convolutional alignment module simplifies the attention mechanism, making it more effective and easier to implement while achieving superior performance.

Handwritten English Word Recognition Using a Deep Learning Based Object Detection Architecture

DOI: 10.1007/s11042-021-11425-7

Main Idea and Contribution

The paper addresses the challenge of recognizing handwritten English words using a deep learning model based on object detection. The authors propose a technique leveraging the YOLOv3 object detection model for sequential character detection and identification, aiming to minimize training data requirements and computational cost. Their approach is lexicon-free, allowing the recognition of words not present in the training set or dictionary.

Dataset(s) Used

- IAM Dataset

Model(s) Used

YOLOv3: A single-stage object detection model adapted for recognizing and localizing characters within handwritten word images.

Procedure/Configuration or Test

- The YOLOv3 model was modified to handle sequential character detection and recognition within handwritten words.
- Training involved using a limited dataset of 1200 word images.
- The model was tested on a larger test set from the IAM dataset.
- The model avoids preprocessing steps such as skew or slant correction.
- It performs character-level segmentation and recognition simultaneously.
- The approach is lexicon-free, making it adaptable to out-of-dictionary words and other Latin script languages.

Performance Metrics

- **Word Error Rate (WER)**: 29.21%
- **Character Error Rate (CER)**: 9.53%

Key Learning/Findings

- A lexicon-free model using YOLOv3 can achieve competitive performance with significantly less training data.

- The proposed method does not require extensive preprocessing and is capable of recognizing words not present in the training set.
- The model's efficiency in training and testing suggests potential for broader applications beyond the IAM dataset, including other Latin script languages**.**

Title/Author/Date	TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models / Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, Furu Wei / September 6, 2022
Conceptual Framework	Introduces a novel OCR framework based on transformers, leveraging pre-trained models for text recognition.
Research Question/ Hypothesis	How can transformer-based pre-trained models improve the performance and generalization of OCR tasks?
Datasets	● IAM Dataset ● Synthesized handwritten text line images by the TRDG2 , an open-source text recognition data generator ● IIIT-HWS dataset ● SROIE dataset
Methodology	Utilizes transformer architecture, pre-training on large-scale text data, followed by fine-tuning on specific OCR tasks.
Analysis & result	The transformer-based approach with pre-training is highly effective for OCR tasks, providing significant improvements over traditional methods.
Conclusions	The TrOCR model demonstrates that pre-trained Transformer-based models can significantly improve OCR performance across different text recognition tasks. The approach is simple, effective, and does not rely on CNN backbones or external language models, making it easier to implement and maintain.
Implications for Future Research	Future work can explore the extension of this approach to other languages and more complex text recognition scenarios.

Title/Author/Date	An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition / Baoguang Shi, Xiang Bai, Cong Yao / 2015
Conceptual Framework	Proposes an end-to-end trainable neural network combining CNN and RNN for sequence recognition tasks, particularly for scene text recognition.
Research Question/ Hypothesis	Can an end-to-end trainable neural network improve the accuracy and efficiency of scene text recognition?
Datasets	● IIIT-5k ● Street View Text ● ICDAR 2003 ● ICDAR 2013
Methodology	Combines convolutional layers for feature extraction with recurrent layers for sequence modeling, followed by a transcription layer for label sequence prediction.
Analysis & result	Demonstrated superior performance on standard benchmarks, achieving state-of-the-art results in both lexicon-free and lexicon-based recognition tasks.
Conclusions	The CRNN architecture is highly effective for sequence recognition tasks, providing significant improvements in accuracy and model compactness
Implications for Future Research	Future research can extend this model to other sequence recognition tasks and explore further optimizations in model architecture and training.

Title/Author/Date	Decoupled Attention Network for Text Recognition/Tianwei Wang, Yuanzhi Zhu, Lianwen Jin, Canjie Luo, Xiaoxue Chen, Yaqiang Wu, Qianying Wang, Mingxiang Cai/2019
Conceptual Framework	Proposes a decoupled attention network (DAN) that separates the alignment and decoding processes for text recognition, aiming to improve accuracy and robustness.
Research Question/ Hypothesis	Can a decoupled attention network improve the performance and robustness of text recognition tasks, particularly for long text and text with disturbances?
Datasets	● I AM ● RIMES
Methodology	The methodology involves a feature encoder, a convolutional alignment module (CAM), and a decoupled text decoder. The CAM replaces the traditional score-based alignment with a fully convolutional network, generating attention maps.
Analysis & result	DAN showed state-of-the-art performance in handwritten and scene text recognition tasks, outperforming traditional attention mechanisms.
Conclusions	The decoupled attention network effectively reduces misalignment errors and improves the robustness and flexibility of text recognition systems.
Implications for Future Research	Future research could explore further improvements in alignment techniques and extend the DAN framework to other text recognition applications and datasets.

Title/Author/Date	Full Page Handwriting Recognition via Image to Sequence Extraction / Sumeet S. Singh, Sergey Karaye / 2021
Conceptual Framework	Proposes a framework to handle the structure of handwritten documents by identifying and classifying regions as text, math, drawing, tables, and deleted text, extracting text in a natural reading order while ignoring untranscribed regions.
Research Question/ Hypothesis	Can a model be trained to effectively transcribe handwritten text from complex full-page documents while ignoring non-text artifacts and untranscribed regions?
Datasets	● IAM dataset ● Free Form Answers dataset
Methodology	Combines a CNN (ResNet) encoder and a Transformer decoder for image to sequence extraction, handling text of varying orientation, layout, and size.
Analysis & result	Achieved state-of-the-art results on paragraph-level recognition on the IAM dataset and performed better than commercially available HTR APIs on real-world handwritten test answers.
Conclusions	The model is effective in recognizing full pages of handwritten text, including various complex elements, and can be adapted to different datasets with retraining.
Implications for Future Research	Future research can explore further optimizations and applications to other domains requiring text recognition from complex layouts.

Title/Author/Date	Handwritten English Word Recognition Using a Deep Learning Based Object Detection Architecture,Riktim Mondal, Samir Malakar, Elisa h. Barney Smith, Ram Sarkar/2022
Conceptual Framework	The paper addresses the challenge of recognizing unconstrained handwritten words using a deep learning approach. It leverages object detection models to recognize characters in words, enabling recognition without segmenting into individual characters.
Research Question/ Hypothesis	Can an object detection-based CNN architecture effectively recognize handwritten words in a lexicon-free approach?
Datasets	● IAM Dataset
Methodology	Uses YOLOv3 architecture for character detection and recognition. The model is tailored to fit the specific problem of handwritten word recognition.
Analysis & result	Achieves satisfactory word recognition results with significantly fewer training samples compared to state-of-the-art methods. The model is effective even with lexicon-free approaches.
Conclusions	Demonstrates the potential of object detection models for handwritten word recognition.
Implications for Future Research	Future research can further refine these models for better accuracy and efficiency.

Project Solution Proposals:

Model Proposed

To address the problem of handwritten document conversion, we propose a two-part model comprising handwriting detection/segmentation and handwriting recognition components. This pipeline leverages advanced techniques in computer vision and neural network architectures to achieve high accuracy in digitizing handwritten text.

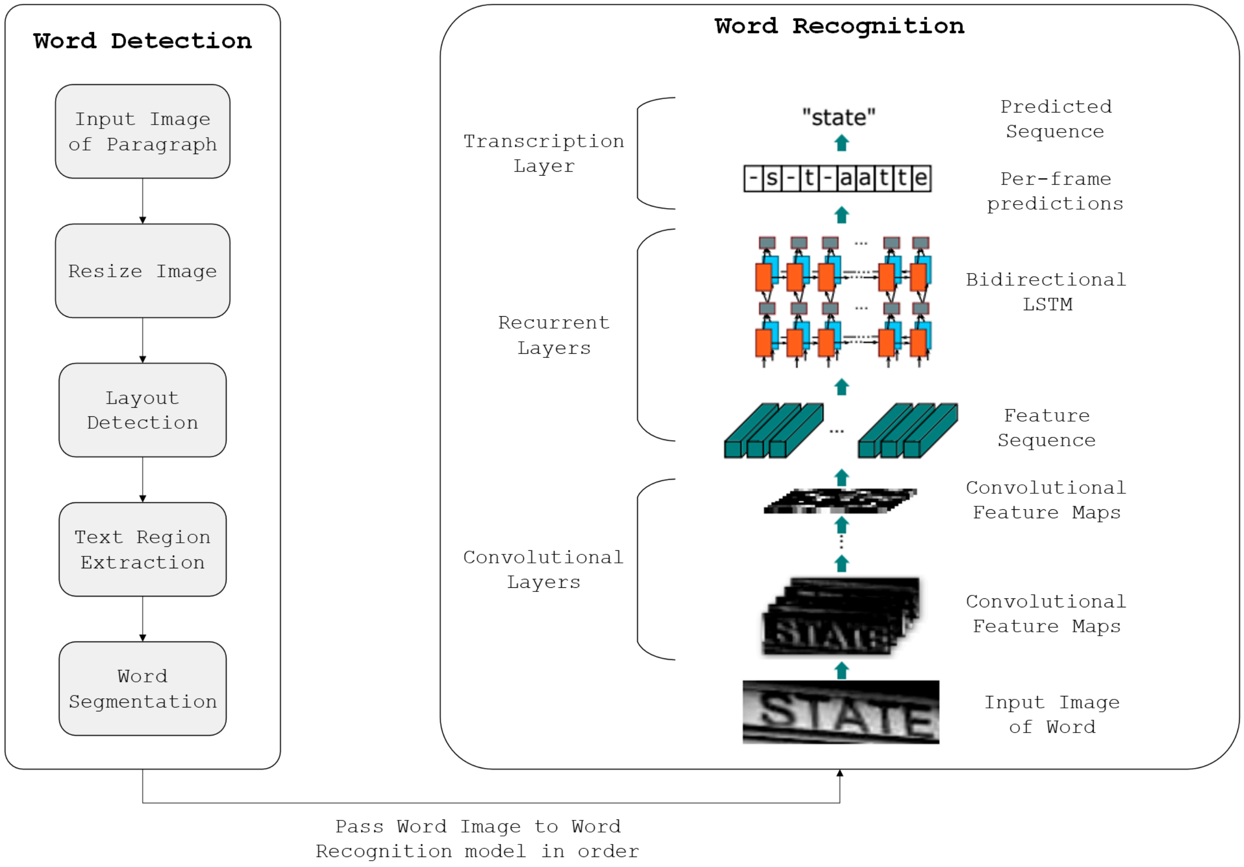
1. **Handwritten Text Detection and Segmentation**
    - **Layout Parsers**: The first part of the model focuses on detecting and segmenting handwritten text from scanned document images. We will use layout parsers to identify and isolate individual characters or words from the document. Layout parsers are designed to analyze the structural layout of documents, making them well-suited for identifying text regions, even in complex layouts. This step involves detecting text lines, segmenting them into words or characters, and preparing them for the recognition phase.
    - **Technique**: Layout parsers utilize computer vision techniques, including convolutional neural networks (CNNs) and transformers, to process the document images and accurately segment the text. By leveraging these powerful models, the layout parsers can handle various challenges, such as overlapping characters and inconsistent spacing, ensuring that the text is properly segmented for further processing.
2. **Handwritten Text Recognition**
    - **CRNN (Convolutional Recurrent Neural Network)**: The second part of the model is responsible for recognizing and transcribing the segmented handwritten text into a digital format. We will use a CRNN, which combines the strengths of CNNs and recurrent neural networks (RNNs). The CRNN first applies convolutional layers to extract features from the segmented text images. These features are then processed by recurrent layers, which capture the sequential nature of handwriting, allowing the model to generate accurate transcriptions.
    - **Language Model**: To further enhance recognition accuracy, we will integrate a language model into the CRNN. The language model will provide contextual information, helping the CRNN to produce coherent and contextually appropriate transcriptions. By leveraging the language model, the system can correct recognition errors and handle ambiguous characters more effectively.

Evaluation Metrics

To assess the performance and effectiveness of our handwritten document conversion model, we will use a combination of evaluation metrics that focus on different aspects of the process, from segmentation to recognition accuracy. The key metrics include:

1. **Character Error Rate (CER):** CER is the ratio of the number of character-level errors (insertions, deletions, and substitutions) to the total number of characters in the reference text.
   $CER\ =\ Insertions\ +\ Deletions\ +\ Substitutions/Total\ Characters\ in\ Reference\ Text$
2. **Word Error Rate (WER):** WER measures the number of word-level errors divided by the total number of words in the reference text. It considers insertions, deletions, and substitutions at the word level.
   $WER\ = Insertions\ +\ \ Deletions\ \ +\ \ Substitutions/\ Total\ Words\ in\ Reference\ Text$
3. **Precision, Recall, and F1-Score**
   - **Precision**: The ratio of correctly predicted text segments to the total predicted segments. It measures the model's ability to avoid false positives.
     $Precision\ \ =\ \ True\ Positives\ /\ True\ Positives\ \ +\ \ False\ Negatives$
   - **Recall**: The ratio of correctly predicted text segments to the total actual text segments. It measures the model's ability to capture all relevant instances.
     $Recall\ \ =\ \ True\ Positive\ /\ True\ Positive\ \ +\ \ False\ Negatives$
   - **F1-Score**: The harmonic mean of precision and recall, providing a single metric that balances both aspects.
     $F1\ Score\ \ =\ \ 2\ *\ \ (Precision\ \ *\ \ Recall)/(Precision\ \ +\ \ Recall)$

Architecture of Proposed System



References:

[1] Li, M., Lv, T., Cui, L., Lu, Y., Florêncio, D. A. F., Zhang, C., Li, Z., & Wei, F. (2021). TrOCR: Transformer-based optical character recognition with pre-trained models. *CoRR*, *abs/2109.10282*. https://arxiv.org/abs/2109.10282

[2] Shi, B., Bai, X., & Yao, C. (2015). An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *CoRR*, *abs/1507.05717*. http://arxiv.org/abs/1507.05717

[3] Singh, S. S., & Karayev, S. (2021). Full page handwriting recognition via image to sequence extraction. *CoRR*, *abs/2103.06450*. https://arxiv.org/abs/2103.06450

[4] Wang, T., Zhu, Y., Jin, L., Luo, C., Chen, X., Wu, Y., Wang, Q., & Cai, M. (2019). Decoupled attention network for text recognition. *CoRR*, *abs/1912.10205*. http://arxiv.org/abs/1912.10205

[5] Mondal, R., Malakar, S., Barney Smith, E. H., & Sarkar, R. (2022). Handwritten English word recognition using a deep learning-based object detection architecture. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-021-11425-7